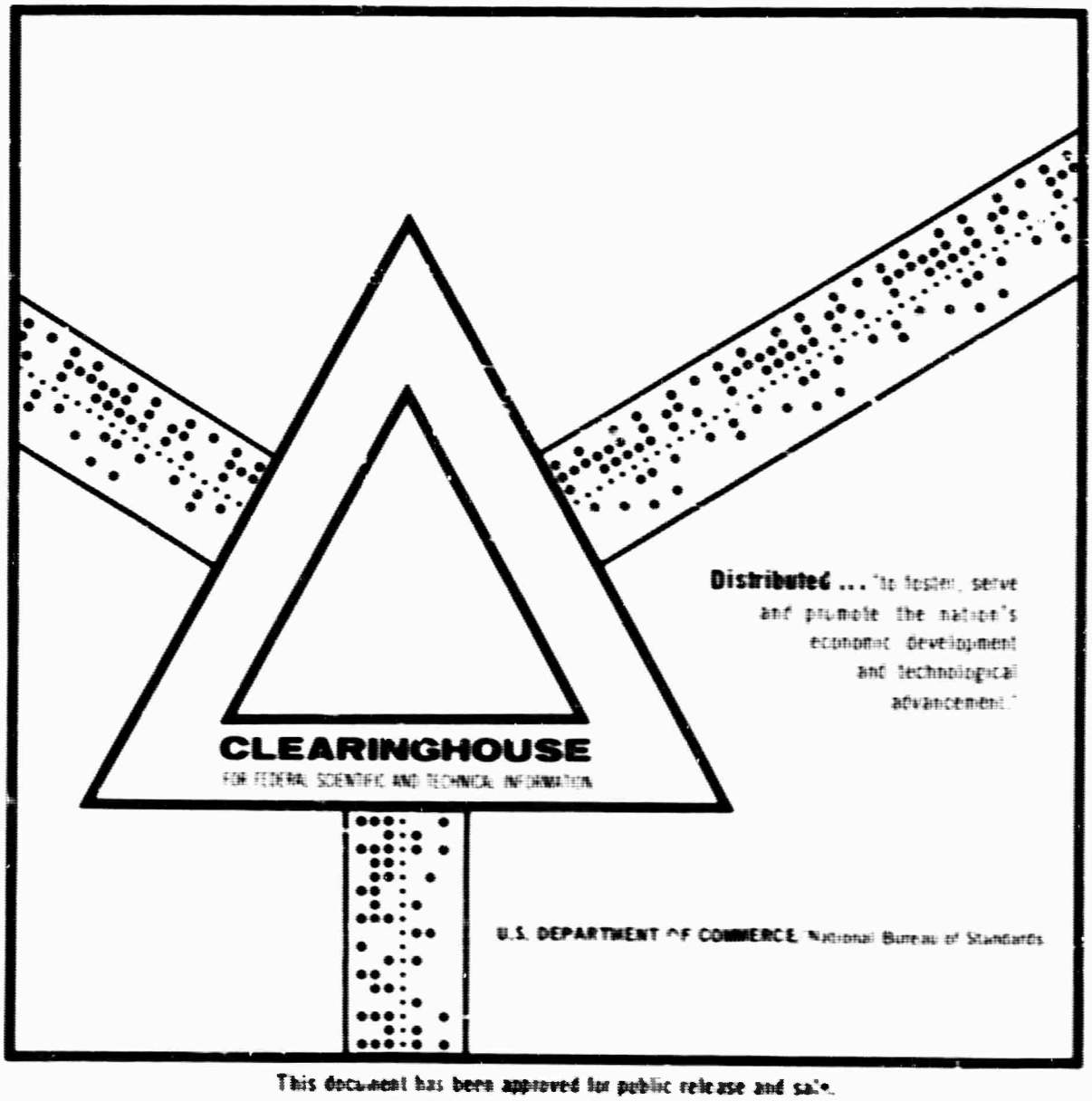
STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS II, AND SOME REMARKS ON THE USE OF ACOUSTIC DATA IN SCHEMES FOR MACHINE RECOGNITION OF SPEECH

Kenneth N. Stevens

Bolt Beranek and Newman, Incorporated
Cambridge, Massachusetts

15 August 1969



DISCLAIMER NOTICE

THIS DOCUMENT IS THE BEST QUALITY AVAILABLE.

COPY FURNISHED CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

BOLT BERANEK AND NEWMAN INC

CONSULTING · DEVELOPMENT · RESEARCH

AFCRL-69-0339

15 August 1969

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS II, AND SOME REMARKS ON THE USE OF ACOUSTIC DATA IN SCHEMES FOR MACHINE RECOGNITION OF SPEECH

by

Kenneth N. Stevens

Scientific Report No. 12
Contract No. F19628-68-C-0125
Project No. 8668
Contract Monitor: Hans Zschirnt, Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts 01730

This research was supported by the Advanced Research Projects Agency under ARPA Order No. 627.

AFCRL-69-0339 15 August 1969

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS II, AND SOME REMARKS ON THE USE OF ACOUSTIC DATA IN SCHEMES FOR MACHINE RECOGNITION OF SPEECH

by

Kenneth N. Stevens

BOLT BERANEK AND NEWMAN INC.
50 Moulton Street
Cambridge, Massachusetts 02138

Scientific Report No. 12
Contract No. F19628-68-C-0125
Project No. 8668
Contract Monitor: Hans Bachlund, Data Sciences Laboratory

Prepared for:

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts 01730

This research was supported by the Advanced Research Projects Agency under ARPA Order No. 627.

ABSTRACT

The acoustic properties of a number of different speech sounds as they appear in several phonetic contexts are described. This report supplements an earlier report on the same topic and presents data for stop and nasal consonants in prestressed position, for the timing of vowels, and for acoustic events following stressed vowels. The aims of this survey are to provide an indication of the kinds of acoustic attributes that should be extracted from the speech signal in a potential scheme for machine recognition of speech. Also included is a discussion of the roles that must be played by acoustic data and by linguistic constraints in schemes for automatic speech recognition.

TABLE OF CONTENTS

LIST OF FIGURES

page

LIST OF TABLES

BLANK PAGE

1. INTRODUCTION

An earlier report* presented a survey of the acoustic properties of a number of different speech sounds as they appear in several phonetic contexts. That survey (hereafter referred to as SK I) was prepared primarily for the use of investigators who are interested in developing procedures for machine recognition of speech, since any such procedure must include a component that extracts certain acoustic properties or attributes from the speech signal. The material in SK I is intended to provide an indication of the kinds of acoustic attributes that should be extracted in a recognition scheme.

The data presented in SK I are far from complete and include, primarily, an analysis of stressed vowels and of consonants in prestressed position. The purpose of this supplementary report is to discuss the acoustic properties of speech sounds in other phonetic environments, particularly when the sounds occur after a stressed vowel.

The point of view in this study is that a given "speech sound" is characterized by several features or properties which can be used to categorize all speech sounds into natural classes depending on their manner and place of articulation. Frequently, the invariant attribute or property that characterizes a natural class or a feature is an articulatory position, posture, or maneuver. The specific acoustic attributes associated with a segment that possesses a given feature may depend to some extent on the features of that

---

*K.N. Stevens and Mary M. Klatt, "Study of Acoustic Properties of Speech Sounds," Report No. AFCRL-68-0446, Air Force Cambridge Research Laboratories, Bedford, Mass. (Aug. 1968). (Also, Bolt Beranek and Newman Inc., Report No. 1669, Cambridge, Mass.)

segment and on the features of adjacent segments. For example, the acoustic attribute that characterizes a coronal consonant (i.e., a consonant produced with the tongue tip) may depend on whether the consonant is a stop, a fricative, or a nasal, and may also depend, in some cases, on the features of the vowel or other segment that follows the consonant. An acoustic description of a feature must, therefore, include data for the feature as it occurs in various contexts.

This supplementary report also includes some remarks (Sec. 4) concerning the problem of automatic speech recognition. These remarks suggest some reasons why automatic recognition of speech can be expected to have only limited success and point out the kinds of "knowledge" a speech recognizer must possess in order to interpret properly the acoustic events and identify the speech units.

The data presented in SK I, as well as in this supplementary report, were based primarily on analysis of a series of utterances (including nonsense syllables) and isolated words produced by three speakers. All of these utterances were processed by a bank of 19 band-pass filters whose rectified and smoothed outputs were sampled, quantized, and printed out. Spectrograms of the recorded material were also produced. The details of the analysis procedures are described in SK I.

2. FURTHER DATA ON CONSONANTS IN PRESTRESSED POSITION

2.1 Stop Consonants

Report SK I presented some samples of data that indicated the acoustic attributes that (1) identify prestressed stop consonants as a class, (2) distinguish between voiced and voiceless stop consonants, and (3) identify place of articulation of a stop consonant as a labial /b, p/, a dental /d, t/, or a velar /g, k/. This Section gives further data (particularly with regard to place of articulation) on stop-consonant properties. Furthermore, attributes of the affricates /č/ and /ǰ/ are discussed, and additional detail regarding the characteristics of stop consonants in initial consonant clusters are presented. It should be noted, incidentally, that stop consonants participate in most of the allowed initial consonant clusters in English.

The average duration of the stop gap for an initial stop consonant (preceded by an unstressed schwa) is in the range 110–130 msec (see SK I, Table VII, p. 42).* These durations apply both to initial single consonants and to stops in initial position in clusters (as in /tr/, /bl/, etc.). For individual utterances, these prestressed stop-gap durations may be as short as 60 msec or as long as 150 msec.

The aspiration interval following release of an initial consonant always identifies a voiceless stop as opposed to a voiced stop. As noted in SK I, this aspiration interval is in the range 50–100 msec for initial, single voiceless stops. The aspiration duration

---

*As noted elsewhere in SK I, this duration can become much shorter for stop consonants in other phonetic environments.

is usually at the upper end of this range (and sometimes even greater) when the voiceless stop is the initial element in a consonant cluster (such as /kr/, /tw/). Generally, when a stop consonant follows an /s/ in an initial cluster, there is no aspiration following the release of the stop. There may be a brief frication interval having a duration less than 10 msec for the labial stop and 20–30 msec for the /t/ and /k/. (These attributes of stop consonants in consonant clusters can be observed in SR I, Fig. 22, p. 67.)

Place of articulation for initial stop consonants is determined by acoustic events in the time interval up to 50 msec after the consonantal release. For voiced stops (or voiceless stops following /s/), this interval usually consists of (1) an initial "transient" associated with the sudden pressure release and a brief frication noise and (2) a voiced interval in which the acoustic spectrum is changing (i.e., the formants are undergoing transitions) as the articulators move from the consonant to the vowel configuration. For voiceless stops, there is the same kind of initial transient, followed by an interval of aspiration.

2.1.1 Labial stop consonants

Following release of the consonant /b/ there is only weak (or nonexistent) frication noise, and the transient interval is very brief (less than 10 msec). Several examples of spectra sampled about 10 msec after release of the consonant /b/ in various environments are shown in Fig. 1. (Since spectra were sampled only every 10 msec, the location of the sampled spectrum relative to the instant of release cannot be determined precisely, but it is probably always in the range 10 ±5 msec.) Figure 1 also shows spectra

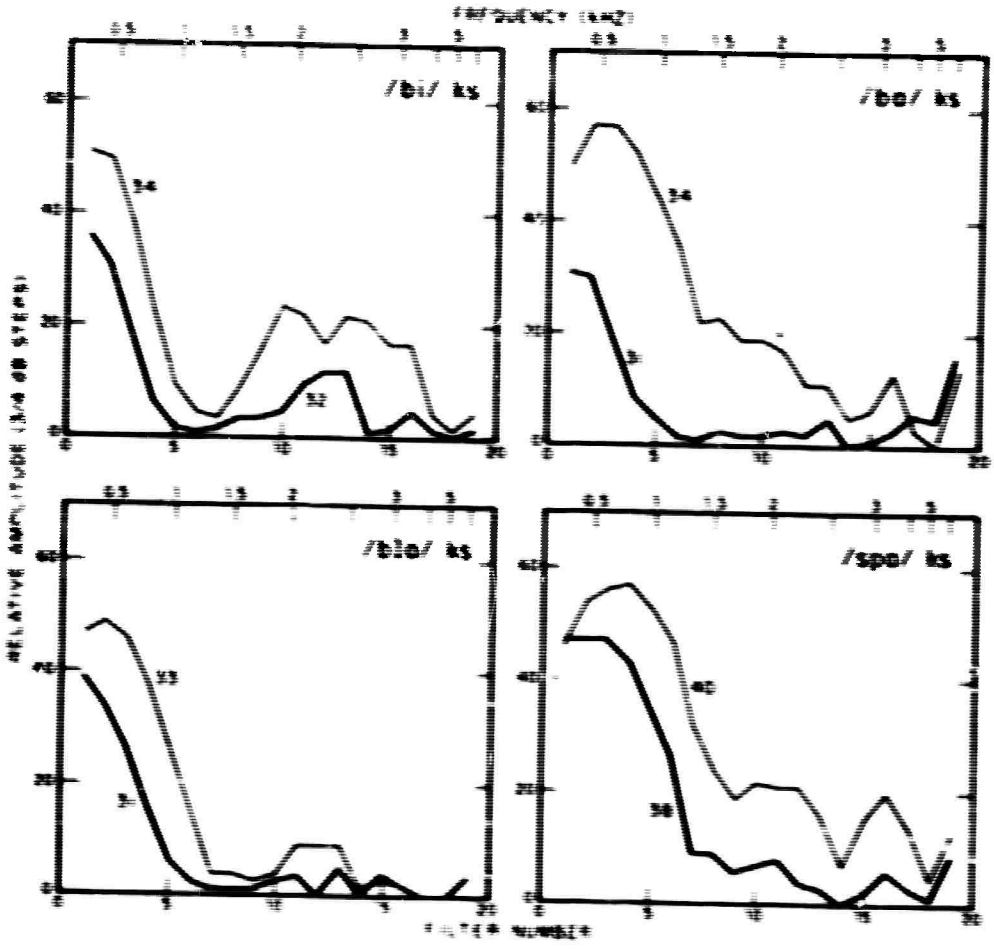


FIG.1 Spectra sampled within about 10 msec of release of unaspirated labial stop consonants (heavy lines) and about 20 msec later during vowel transition (light lines). Spectra are obtained from 19-channel filter bank described in SR 1; curves are identified by sample numbers representing 10-msec intervals.

sampled about 20 msec later, during the transition into the following sonorant (although in some cases the transition is essentially completed by this time). In all examples, the spectrum at high frequencies (above, say, 3500 Hz) in the initial 10 msec is relatively weak compared to the spectrum amplitude in the following vowel; and the components with the greatest amplitude, which are in the low-frequency range, are a consequence of the glottal source rather than the acoustic transient at the lips.

Additionally, Fig. 1 displays spectra sampled at comparable intervals of time following release of the /p/ in the cluster /rpe/. These spectra have characteristics very similar to those of the /t/ spectra.

Spectra of initial stop consonant /p/ sampled about 10 msec after release are shown in Fig. 2; also displayed are spectra sampled about 20 msec after onset of voicing of the vowel. Since there is an interval of aspiration, the vowel spectra are usually sampled 50-100 msec after the consonantal release. As in the case of the /t/ spectra, the initial transient interval for the /p/ is generally weak compared with the following vowel (except possibly at very high frequencies, above about 4500 Hz). The spectra show no pronounced energy peaks, which indicates that none of the formants are strongly excited.

2.1.2 Postdental stop consonants

Immediately following the release of the consonant /d/ in prestressed position, there is a brief interval of frication noise of about 20-msec duration. (This attribute of the consonant /d/ can be observed on the spectrograms in SR 1, Fig. 8, p. 38.) The spectral properties of this initial energy burst are shown in Fig. 3. The

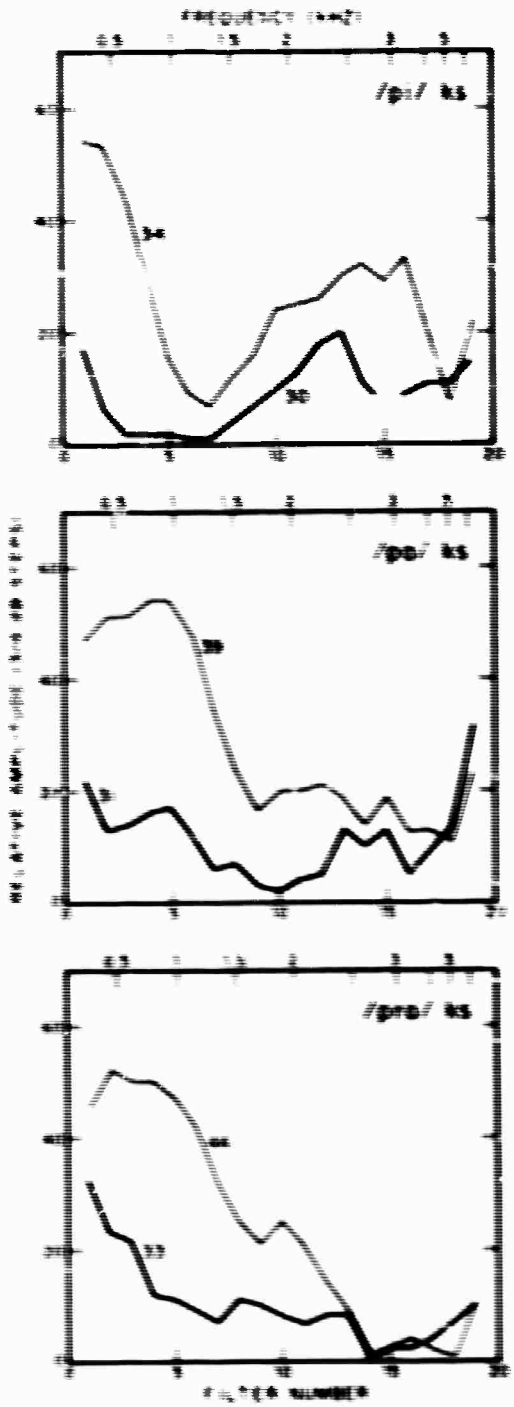


FIG. 2 Spectra sampled within about 10 msec of release of aspirated labial stop consonants (heavy lines) and about 20 msec after onset of voicing of following vowel (light lines). [See legend of fig. 1.]

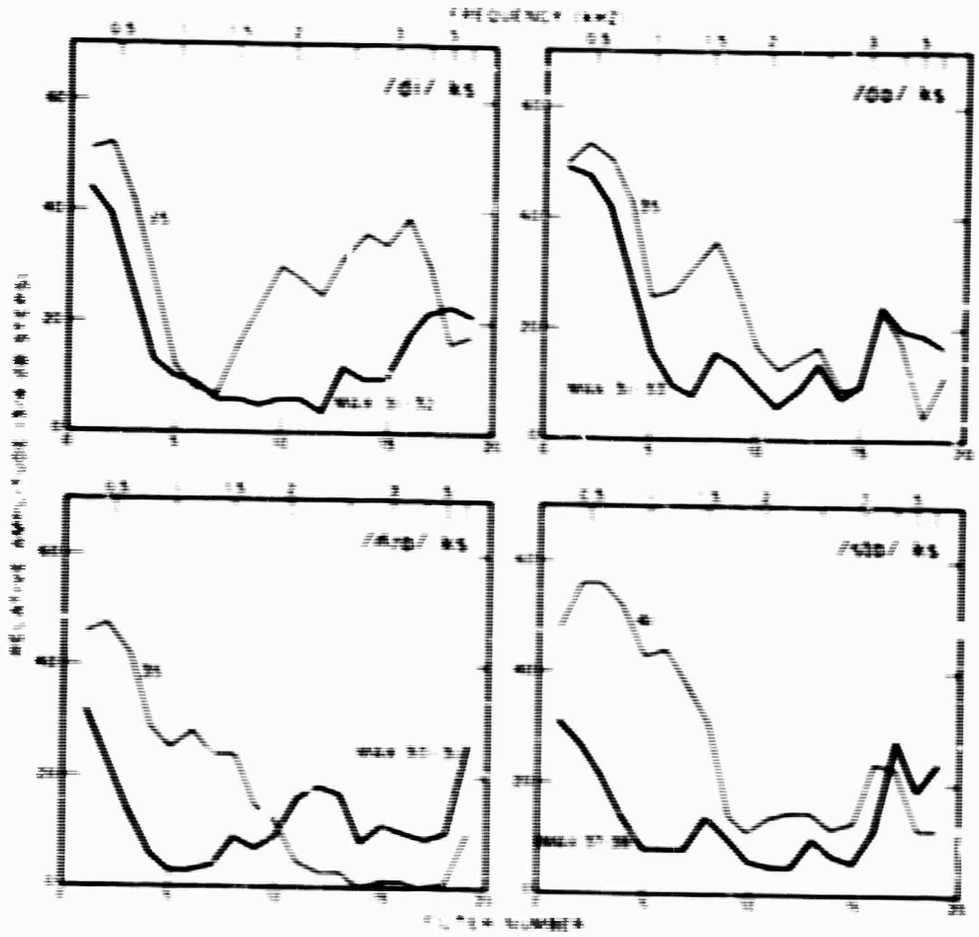


FIG.3 Spectra sampled within about 20 msec of release of unaspirated postdental stop consonants (heavy lines) and about 20 msec later during vowel transition (light lines). The initial spectra represent maximum levels in each filter over the frication interval, as indicated by the sample numbers. [See legend of Fig. 1.]

value plotted for each filter is simply the maximum output of that filter during the frication noise interval (usually 20 msec, or two sample periods). Figure 3 also displays the spectra sampled 20-30 msec after the burst, during the transition into the vowel. Spectra of the stop consonant in /sto/ are also shown in Fig. 3, since the characteristics of the unaspirated /t/ in this syllable are similar to those for /d/ (except that there may be less low-frequency energy in the unaspirated /t/, since voicing may not be developed fully in the burst interval).

The distinguishing characteristics of the spectra associated with /di/, /do/, and /sto/ are the energy concentration in the burst at high frequencies (above 4000 Hz) and the lack of major spectral peaks below this frequency (except for the peak corresponding to the first formant). The amplitude of the high-frequency spectral peak for the burst is greater than the spectrum amplitude in the same frequency region for the following vowel. Thus, the high-frequency energy burst precedes the onset of the major vowel spectral peaks (except for F_1) at lower frequencies.

The spectrum of the /d/ burst in /dro/ is a special case, since /d/ preceding /r/ in English is usually produced with a constriction further back in the mouth. This more posterior constriction gives an initial energy burst with a peak at a lower frequency (about 2300 Hz in this case).

Spectra sampled about 10 msec after the onset of /t/, shown in Fig. 4, are similar in form to those for /d/, except that the high-frequency energy concentrations are stronger, and there is, of course, no low-frequency energy due to voicing.

Again, the amplitude of the high-frequency peak is greater than the spectral amplitude in the following vowel in the frequency range.

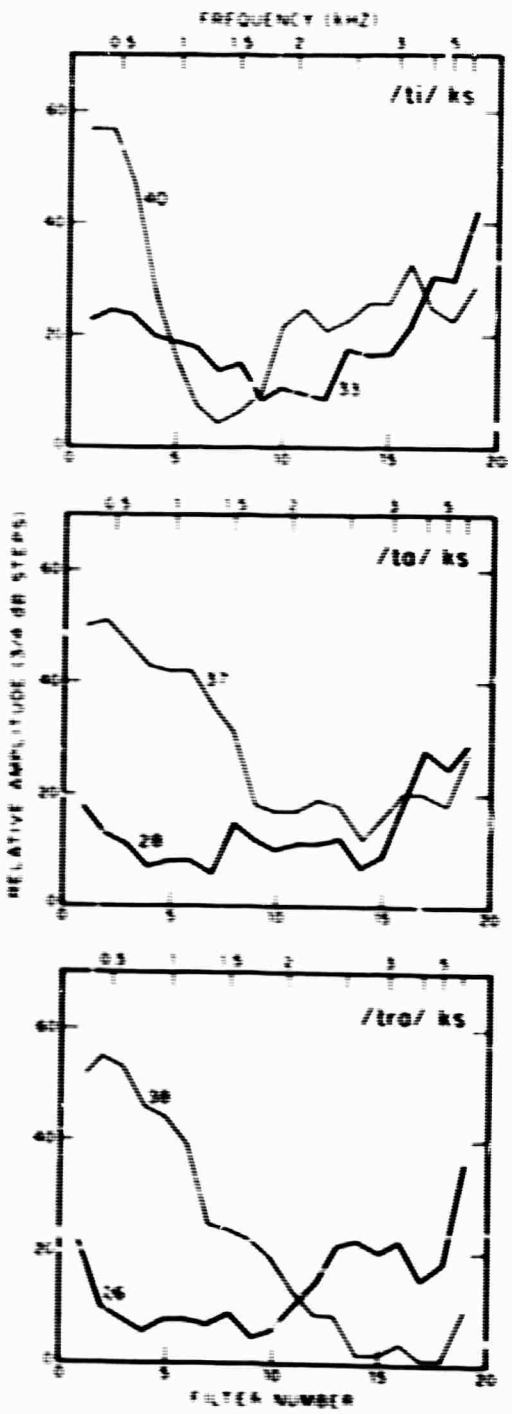


FIG. 4 Spectra sampled within about 10 msec of release of aspirated postdental stop consonants (heavy lines) and about 20 msec after onset of voicing of following vowel (light lines). [See legend of Fig. 1.]

2.1.3 Velar stop consonants

The burst of frication noise following release of the consonant /g/ usually has a duration of about 30 msec. This longer interval of frication noise provides one cue for distinguishing /g/ from /b/, since the frication noise for /b/ is much briefer.*

Spectra for the /g/ burst, displayed in Fig. 5, were obtained by taking the maximum value for each sampled filter output during this 30-msec interval. As before, the spectra sampled 20–30 msec after the burst are shown, as are spectra corresponding to the unaspirated stop consonant in the syllable /sk /.

In order to interpret these data, it is necessary to make a distinction between /g/ preceding a back vowel and /g/ preceding a front vowel. In the environment of a front vowel (/i/ in this case), the /g/ burst has major energy peaks in the high-frequency region (at about 2400 Hz and 3300 Hz in this example). These peaks are comparable in amplitude to the vowel spectrum in this frequency region (corresponding to F_2 and F_3 for a front vowel). When /g/ precedes a back vowel, the major spectral peak in the burst is at a lower frequency. This peak is in the vicinity of the second formant of the vowel and is again comparable in amplitude to the vowel spectrum amplitude in that frequency range.

---

*The existence of a burst of frication noise for /g/ is often not easy to see from the display of the outputs from the 19-channel filter bank. Frication noise cannot be distinguished from voiced excitation, since the averaging times of the smoothing filters are too great (see SK I, Fig. 2, p. 8). However, the frication interval can be seen easily on the spectrogram (which is produced with a much shorter averaging time – of the order of 3 msec).

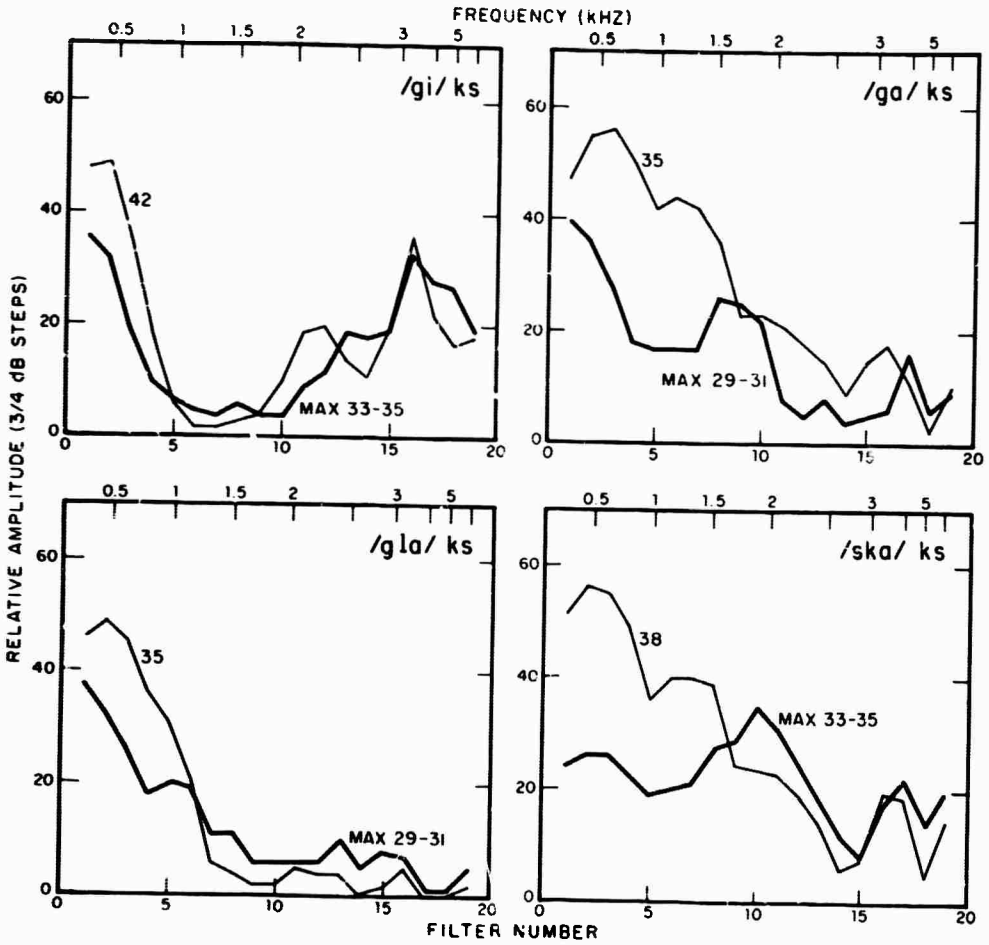


FIG.5 Same as Fig. 3 except for unaspirated velar stop consonants.

Similar remarks may be made with regard to the spectra of /k/ shown in Fig. 6. Here, the spectra at the onset were sampled about 10 msec after the release of the stop consonant. The spectrum of the /k/ burst has a high-frequency peak (about 2600 Hz) when it precedes a front vowel and a peak at lower frequencies when it precedes a back vowel (or when it precedes /r/, /l/, or /w/).

2.1.4 Affricate consonants

There are two affricate consonants in English – voiceless /$\overset{v}{c}$/ and voiced /$\overset{v}{j}$/ – both of which can occur in either initial or final position in a syllable. The duration of the stop gap preceding the release of these consonants in prestressed position is similar to that for other stop consonants (in the range 90–120 msec for the utterances examined in this study). There is a long frication interval following the release; the average duration of frication for /$\overset{v}{c}$/ is about 100 msec and for /$\overset{v}{j}$/ about 70 msec for the three speakers used in this study.

The spectrum of the /$\overset{v}{c}$/ during the frication interval is very similar to that of the /$\overset{v}{s}$/, and the /$\overset{v}{j}$/ and /$\overset{v}{z}$/ also have comparable spectra. Figure 7 shows the spectra of /$\overset{v}{c}$/ preceding each of three vowels and of /$\overset{v}{j}$/ preceding the vowel /a/, all produced by one speaker. The spectrum peaks at about 2500 Hz and 3300 Hz, corresponding to F_3 and F_4, and are evident in all of these spectra. The effect of voicing can be seen at the low-frequency end of the /$\overset{v}{j}$/ spectrum.

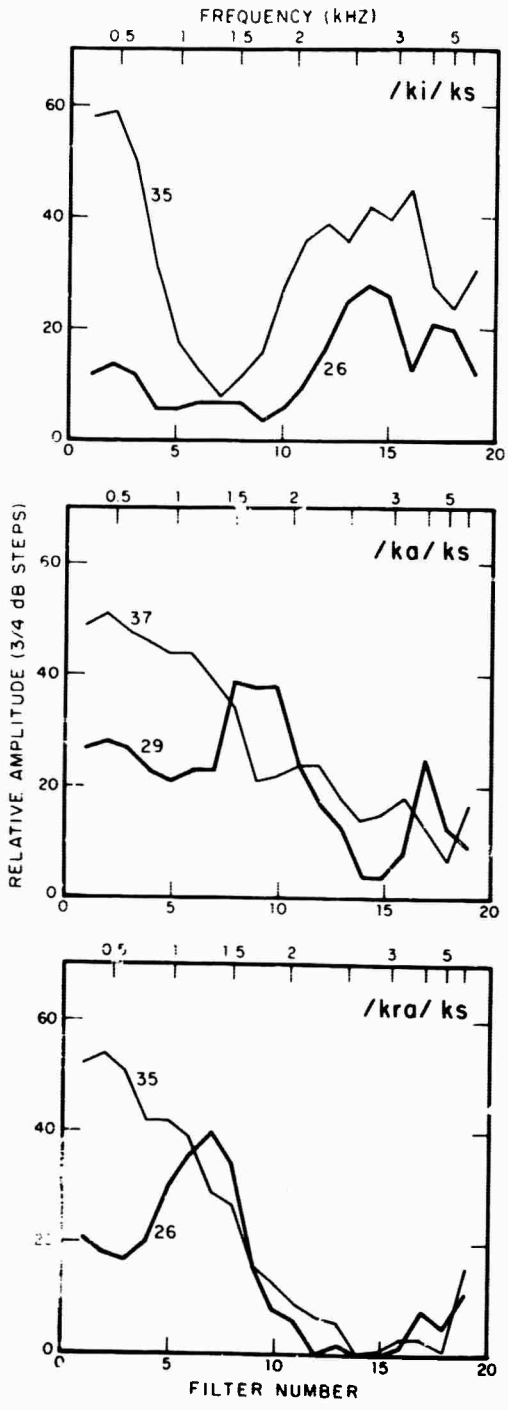


FIG.6 Same as Fig. 4 except for aspirated velar stop consonants.

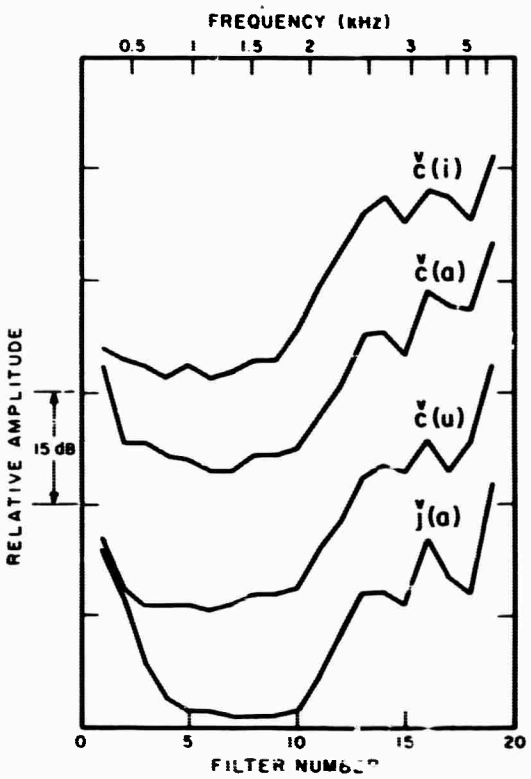


FIG.7 Spectra of three examples of /v̆c/ and one example of /v̆j/ in the vowel environments shown, sampled during the frication intervals. Speaker KS.

2.2 Nasal Consonants

It was observed in SK I that nasal consonants as a class are characterized by an interval of time in which the spectrum remains relatively fixed, followed by a rapid change in spectrum as the consonant is released. The spectrum within the closure interval has a major peak in the range 200-300 cps (filters 1 and 2 for the analyzing system used in this study), and the amplitude at high frequencies is relatively low. Nasal consonants are always characterized by a minimum in spectrum amplitude around 800 Hz (filter 4 in the analyzing system used here).

The two nasal consonants that can occur in prestressed position in English can be distinguished from each other on the basis of the rapid changes occurring in the signal at the instant of consonantal release. The nasal murmur preceding release does not shown consistently different characteristics for /m/ as opposed to /n/. (Some indication of the kinds of acoustic attributes that separate /m/ from /n/ has been presented in SK I, Fig. 19, p. 59.) Here, more detailed data for /m/ and /n/ are shown in Figs. 8 and 9, respectively. Each portion of these Figures shows a pair of spectra obtained from the 19-channel filter bank. One of the spectra (the one drawn with heavy lines) is sampled immediately before the consonantal release, and the other is sampled about 20 msec later. [The spectra sampled during rapid changes in the signal are, of course, very much dependent on the characteristics of the analyzing system, particularly the smoothing filters (see SK I, Fig. 2, p. 8)].

When the nasal consonant precedes the vowel /a/ (or, more generally, when it precedes a back vowel), there is a rapid and large jump in spectral energy in the vicinity of 1.7 kHz following release of /n/;

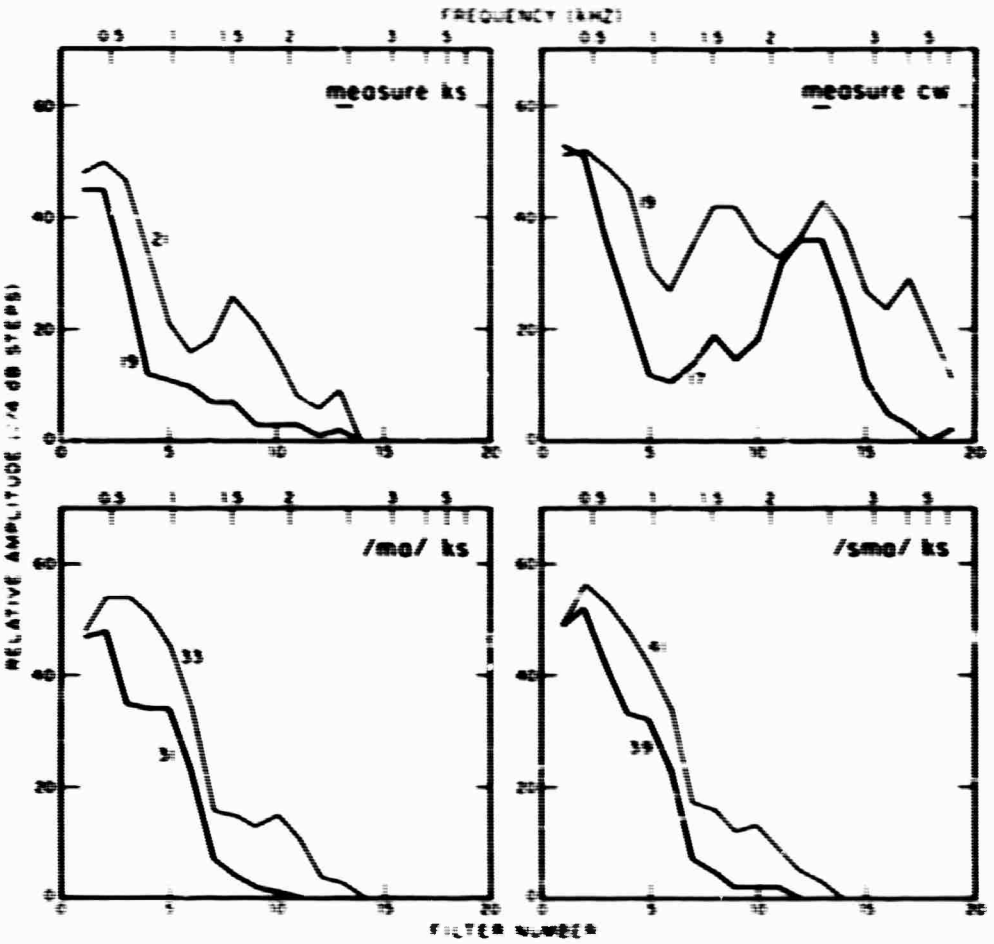


FIG.8 Spectra sampled immediately preceding release of consonant /m/ (heavy lines) and about 20 msec after release into following vowel (light lines).

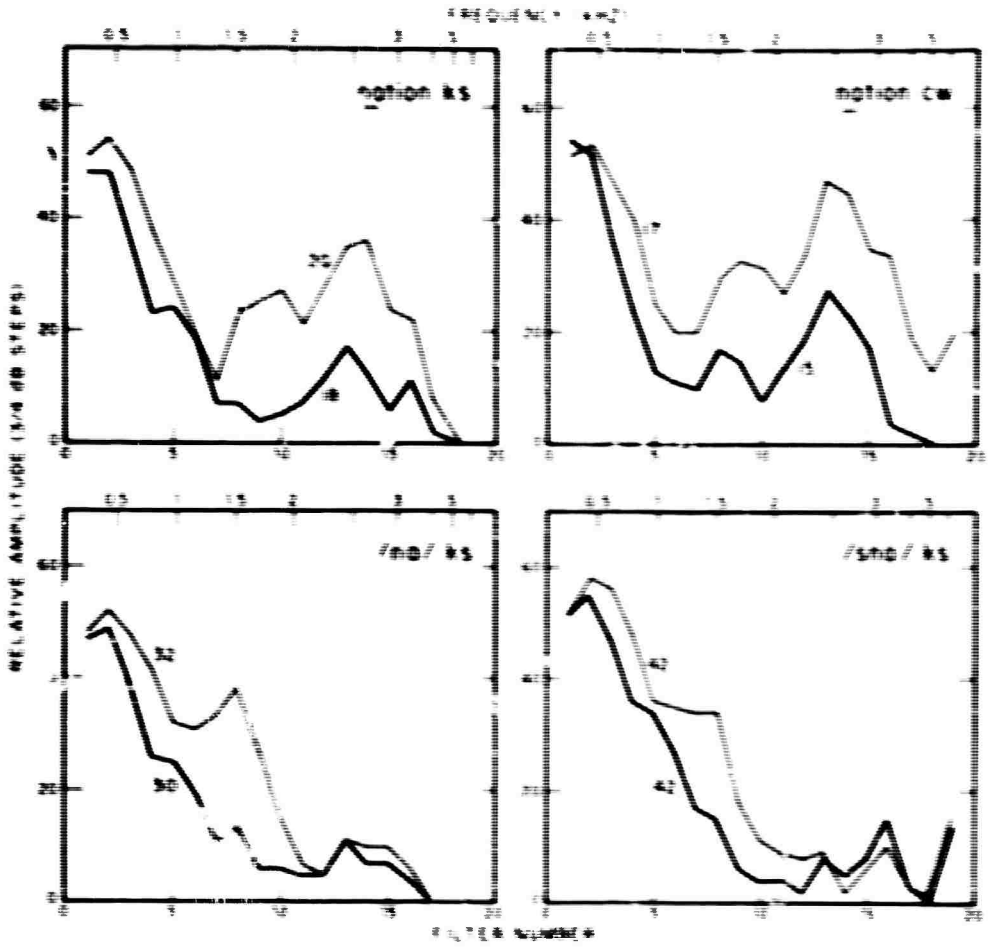


FIG.9 Same as Fig. 8 except consonant is /s/.

this jump in amplitude is much smaller for /n/. When the following vowel is a front vowel (/ɛ/ in the examples shown), there is again a large jump in spectral energy near the high-frequency edge of the major vowel energy concentrations (around 2.5 kHz in the example shown in Fig. 9) following the /n/ release but a much smaller jump following the /m/ release.

Acoustic characteristics of this kind indicate that the frequency region where the onset properties following a nasal are to be examined depends on the vowel, and there are no measurements in an absolute frequency region that will serve to distinguish /m/ from /n/.

3. FURTHER DATA ON TIMING OF STRESSED VOWELS AND ON ACOUSTIC EVENTS FOLLOWING STRESSED VOWELS

3.1 Introduction

Much of the material presented in CR 1, as well as in Sec. 2 of this report, was concerned with the acoustic properties of stressed vowels and of consonants preceding stressed vowels. Only brief consideration was given to acoustic events occurring at the ends of stressed vowels.

For purposes of discussing these events in poststressed positions, three different kinds of phoneme sequences that can follow stressed vowels are considered.

(1) The stressed vowel occurs in syllable-final position. The only utterances of this kind that were examined in this study were bisyllabic nonsense utterances of the form /ə'CV(C)/ and a few monosyllable words. If consideration is restricted to monosyllable single-morpheme utterances (i.e., omit consideration of plural nouns, past tense for vowels, etc.), then the terminations of vowels in this position can be of two kinds: a single consonant, as in /ə'bi/, and a cluster of two consonants, as in the words held, heart, feed, etc.

It may be of interest to note that there are severe constraints on final consonant clusters in this situation. The clusters can consist of: /l/, /r/, or a nasal followed by a voiced or voiceless stop or a voiceless fricative (milk, hand, hark); /l/ or /r/ followed by a nasal (harm, hilt); a voiceless fricative-stop sequence (diet); a voiceless stop-fricative sequence (lapse); and, rarely, a sequence of two voiceless stops (act), the last of which is always a /t/. Within this set of possibilities there are still further constraints, some of which

are imposed by the preceding vowel: For example, /ts/ is not possible in final position except in a two-morpheme situation; /wosp/ is possible but /woisp/ is not permissible; and for a cluster in which the first element is a nasal, both segments have the same place of articulation, so *lump* is possible but *lumt* is not.

(2) A second class of segment sequences that can follow a stressed vowel is a consonant (or consonant cluster) followed by an unstressed vowel, possibly followed in turn by a final consonant. The unstressed syllable can terminate in one or two nonsonorant consonants (*famous*, *rabid*) or in a sonorant consonant (nasal, or /w, j, r, l/). In many situations, the unstressed vowel /ə/ followed by a sonorant consonant simply reduces to a syllabic sonorant. Examples of words of this type are *legal*, *father*, and *lemon*. In words such as *cocoa* and *baby*, the final syllables could be represented as /əw/ and /əy/, respectively, but in the latter case, the termination reduces simply to /y/ (or /i/).

(3) Finally, a stressed syllable can be followed by a syllable with secondary stress, as in the words *essay* and *seaweed*.

The acoustic events associated with the phonetic segments that follow a stressed vowel can be presented in terms of (1) the effect on the stressed vowel, particularly its duration, (2) the properties of single final consonants, (3) the properties of the components of final consonant clusters, (4) data on poststressed intervocalic consonants, and (5) data on final unstressed syllables.* As

*Some discussion of items (4) and (5) has been given previously in SR 1.

observed elsewhere in this report, the acoustic attributes that result from a phoneme or from a feature are often very much dependent on the context in which the phoneme or feature occurs. This influence of context is particularly strong in segments in poststressed position.

When one examines the acoustic characteristics of a consonant in prestressed position, the primary concern is with events occurring during the constricted interval and immediately following release of the consonantal closure. The release gesture is often called the explosion phase of the consonant, particularly when the consonant is a stop or a nasal. When the consonant *follows* a stressed vowel (or, for that matter, when any vowel precedes the consonant in a syllable), the identity of the consonant may be signaled by acoustic events that immediately precede the closure interval. This interval is often called the implosion phase of the consonant.

In SR I, attention was focused primarily on the constricted interval and on the explosion phase of consonants in prestressed position. This Section of the supplementary report presents some discussion of the characteristics of the implosive phase following a stressed vowel.

3.2 Timing of Events Following Stressed Vowels

The duration and timing of events in the acoustic speech signal provide two kinds of information that may be useful in schemes for automatic speech recognition. First, some features are identified on the basis of a duration measurement. For example,

an initial voiceless stop consonant differs from a voiced stop on the basis of the duration of aspiration preceding the onset of voicing; or, the presence of an initial stop consonant requires a silent interval whose duration exceeds a certain value (probably around 50 msec); or, as noted below, the voicing feature of a poststressed stop or fricative consonant is determined by timing of events in the preceding vowel. Some knowledge of the timing of speech events also helps to establish where certain acoustic characteristics in the signal are likely to occur and hence indicates where specific acoustic measurements on the signal are to be made.

In the kinds of utterances examined in this study, the stressed vowel is either the final vowel or is followed by one other vowel that does not have primary stress. Consequently, comments on timing are restricted to utterances of this type.

As indicated in SK I, pure, nondiphthongized stressed vowels in English can be either long (ɑ, ɔ, æ) or short (ɪ, ɛ, ʌ, ʊ). It might be argued that a long vowel is, in essence, a vowel-vowel combination, although the length of a long vowel is not quite double that of a short vowel. In the environment b--b, the durations of long vowels in single-syllable utterances are, on the average, 320 msec, whereas, on the basis of the data examined in this supplementary study, the durations of short vowels are 180 msec. Other stressed vowels, like /i, e, o, u/, are always followed by sonorant consonants to yield /iy, ey, ow, uw/; sometimes /ɑ/ and /ɔ/ are followed by sonorants to give the diphthongs /ɑy/, /ɑw/, or /ɔy/. These diphthongs and diphthongized vowels are about equal in length to long vowels in the environment b--b. The vowel /ɝ/ can probably also be classified as a long vowel; it is,

in some sense, a distorted or degenerate version of a short vowel (/ɪ/ or /ʌ/) followed by the consonant /r/.

The duration of a stressed vowel or diphthong that is the nucleus of the final syllable in an utterance is influenced by the voicing feature of the final consonant. If the consonant is voiced, the vowel is lengthened; if it is voiceless, the vowel is shortened. The vowel is also lengthened if there is no consonant in final position. Average durations of vowels following the voiced and voiceless obstruents and nasals, as reported by House,* are shown in Table I.

TABLE I. Average vowel durations for symmetrical consonant-vowel-consonant syllables in English. Data for each consonant represent averages over 36 utterances (12 vowels, 3 speakers). [From A.S. House, "On Vowel Duration in English," *J. Acoust. Soc. Amer.* *33*, No. 9, 1174–1178 (1961).]

Consonantal Environment	Vowel Duration (msec)
b	270
d	310
p	150
t	150
m	240
n	260

*See legend for Table I for complete reference.

When a stressed vowel is in the final syllable and is followed by a consonant cluster (consisting of a consonant followed by an obstruent consonant, as in *bond, fault,* etc.), then the voicing feature of the final consonant influences the duration of the vowel-sonorant combination in much the same way that single vowel durations are influenced. The altered duration occurs both on the vowel segment and the sonorant portion. When the final consonant cluster is a sequence of two obstruent consonants, as in *lots* and *sods,* then both of these segments always have the same voicing feature, and this voicing feature again influences vowel duration in the same way. These effects on vowel duration for the few single-syllable words examined in this study are given in Table II. For the syllabic nuclei with sonorant consonants, the vowel-plus-sonorant combinations are somewhat greater than the durations of the single-vowel nuclei, except when the sonorant is /r/.

When a stressed vowel is followed by another syllable that does not have primary stress, the presence of this additional syllable causes a reduction in length of the stressed vowel. The vowel durations tabulated in SK I show marked differences for stressed vowels in bisyllabic words having stress on the first syllable. In fact, it is generally observed that any vowel in the final syllable of an utterance is longer than the same vowel in another position.

If rough estimates of average durations are made from the data in Tables IV and V of SK I (see SK I, pp. 19 and 20), one finds that the average duration of short vowels in bisyllabic words is about 110 msec and that of long vowels (or vowel-plus-sonorant combinations) is about 180 msec. For monosyllabic words, the corresponding averages are about 180 and 320 msec as noted above. Thus, approximately 70 msec of duration is added to stressed short vowels

TABLE II. Durations of vowel and sonorant segments in monosyllabic words terminating in voiced and voiceless consonants (average data for three speakers).

Word	Vowel Duration (msec)	Sonorant Duration (msec)	Vowel and Sonorant (msec)
gaunt	240	50	290
fault	200	60	260
heart	100	80	180
bond	310	140	450
bald	310	110	420
hard	200	130	330
lots	150	–	150
sods	300	–	300

when they appear in syllable-final position; and about 140 msec is added to long vowels in the same situation.

In the two-syllable utterances with stress on the first syllable, the voicing characteristics of an intervocalic consonant following the stressed vowel have only a weak influence on the duration of the stressed vowel. This influence appears to be stronger when the stressed syllable forms a single morpheme (e.g., *seated* vs *seeded*), and the effect is often much smaller for one-morpheme

words (e.g., *rabid* vs *rapid*). Presumably in the latter case, the voicing feature of the consonant is signaled by other acoustic events such as aspiration or vocal-cord vibration during the stop gap. Thus the vowel durations in *seated* and *beaded* (averaged over the three speakers) are 110 and 160 msec, respectively, while the vowel durations in *rapid* and *rabid* are 170 and 200 msec, respectively.

These data on the timing and duration of events within stressed vowels may be summarized as follows. The duration of a stressed vowel, a diphthong, or a vowel-plus-sonorant sequence (excluding nasals) may range from about 80 msec to more than 400 msec (in "normal" speech production), depending on the features of the vowel and on phonetic events following the vowel. Measurements on the vowel spectrum in the region 50–100 msec following release of the initial consonant or consonant cluster can usually serve to make an identification of the place of articulation of the vowel, i.e., the status of the features high, low, back, and rounded. If the vowel interval extends beyond this time, additional measurements in the following time interval can be made to determine whether the vowel is long or short and whether it is a diphthong or is followed by a sonorant consonant. If it is determined that the vowel is followed by an obstruent consonant, then duration measurements on the vowel may be needed to determine whether the consonant is voiced or voiceless.

Unstressed vowels are, of course, generally shorter than stressed vowels, but their durations undergo influences that are similar to those for stressed vowels. Thus, an unstressed vowel that is the nucleus of a final syllable (as in *famous*) is usually longer than an unstressed vowel in an earlier syllable (as in *about*). Likewise, the voicing characteristics of a final consonant

following an unstressed vowel influences the vowel duration. For example, the average unstressed vowel duration in *raucous* is 70 msec (for the three talkers in this study), whereas in *edges* it is 160 msec; but these durations are quite variable from one talker to another.

As noted in SK I, the duration of the constricted interval for a single consonant is always less when the consonant follows a stressed vowel and precedes an unstressed vowel than for a consonant in prestressed position. This interval may be as short as 15–20 msec when the intervocalic consonant is a dental stop or nasal and may be as long as 150 msec for a voiceless fricative consonant in this position.

3.3 Single Poststressed Consonants in Final Position

The acoustic attributes that characterize a consonant in final position are often markedly different from the attributes of the same consonant in prestressed position, particularly with regard to the way voicing and manner of articulation are signaled. The following subsections present brief comments on the various classes of consonants that appear in final position.

3.3.1 Final fricative consonants

The spectra of voiceless fricative consonants are very much the same whether the consonants appear in prestressed position or in poststressed final position. Thus, the spectra shown in Fig. 12

of SK I (see SK I, p. 45) indicate the main features of voiceless fricative consonant spectra for any phonetic environment.

Spectra sampled in the fricative noise portion of a final *voiced* fricative consonant indicate that this segment is voiceless or only weakly voiced throughout most of its length. This lack of voicing (or weak voicing) can be seen in the spectrograms shown in Fig. 13 of SK I (see SK I, p. 47). Examples of spectra sampled in the middle of the consonantal interval for both initial and final voiced fricatives are compared here in Fig. 10.

All of these examples show appreciably less low-frequency energy in the spectra of the final consonants and indicate that there is little or no vocal-cord vibration in the consonantal intervals. There are also some differences in the spectra at high frequencies, suggesting that there may be some differences in tongue position and degree of constriction for the final consonants relative to the initial consonants.

The spectra in the constricted intervals of final voiced fricative consonants are, therefore, very similar to those for voiceless consonants. The difference between a final voiced and voiceless fricative is signaled largely by the time course of the vowel preceding the consonant. Not only is the vowel longer before a voiced fricative (as noted in Sec. 3.2), but the vowel configuration in the few tens of milliseconds prior to the onset of frication noise for the consonant is generally more constricted. This difference is illustrated in Fig. 11, which shows vowel spectra sampled about 70 msec prior to onset of the consonant in the syllables /sɑs/ and /zɑz/. It is evident that the first-formant frequency is much lower in the case of the voiced consonantal environment, with the result that there is much less energy in the vowel

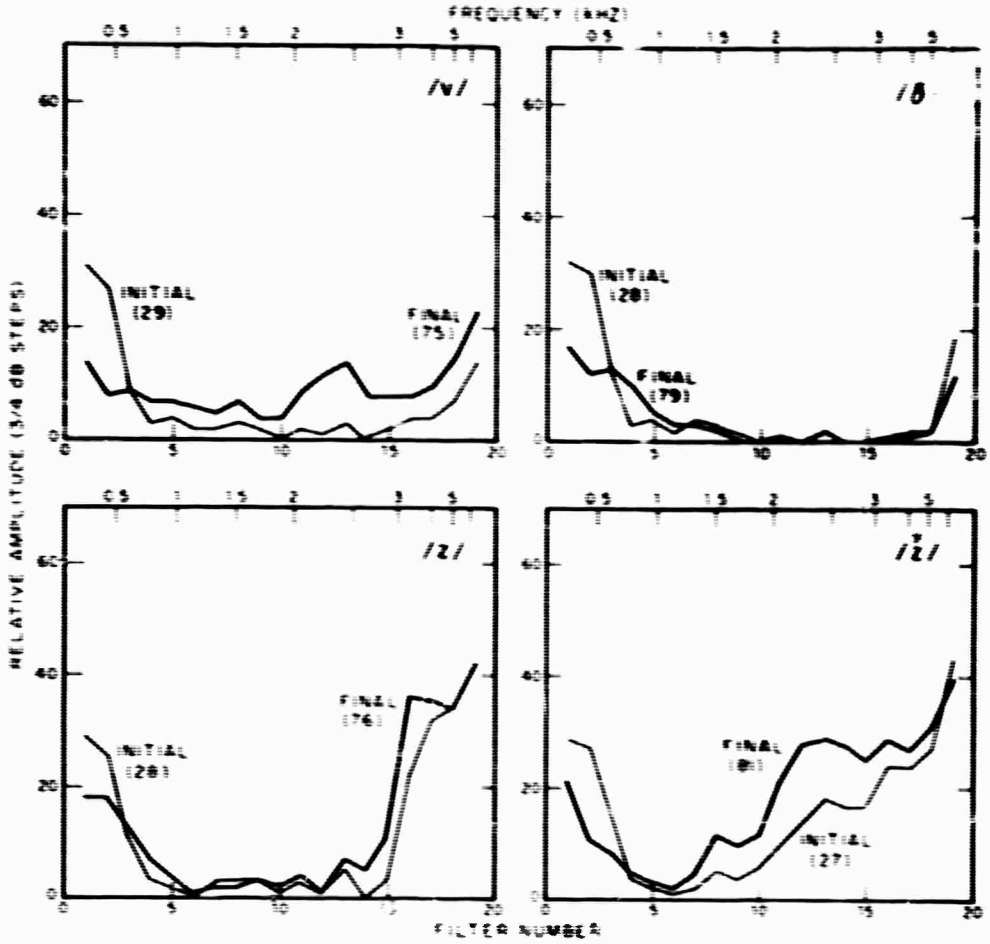


FIG.10 Spectra of voiced fricative consonants sampled during constricted intervals preceding a stressed vowel (light lines) and in terminal position following the same stressed vowel (heavy lines). Sample numbers are identified. Speaker KS.

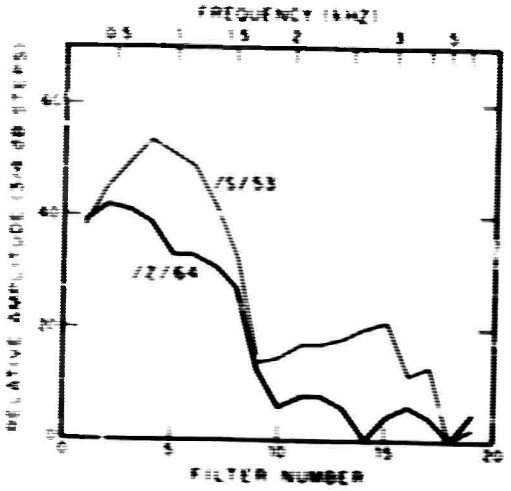


FIG.11 Spectra sampled about 70 msec prior to onset of frication noise in the syllables /cs/ light lines and /oz/ (heavy lines). Speaker KS.

spectrum over the frequency range above the first formant. Similar differences are apparent for other vowels and other final fricative consonants. It might be argued, then, that a final voiced fricative (and, probably also, a final voiced stop) is often characterized by an attribute or feature that lengthens the vowel or, equivalently, delays termination or closure of the vowel.

3.3.2 Final stop consonants

A stop consonant in final position also has an influence on both the duration and the time course of the vowel preceding it, depending on whether the consonant is voiced or voiceless. This influence on the vowel is similar to that of a fricative consonant as described in Sec. 3.3.1.

The place of articulation of a stop consonant in final position is signaled in part by the transition or change in the acoustic spectrum of the vowel in the few tens of milliseconds preceding the consonant implosion. It frequently happens in English that a final stop consonant is not released (i.e., has no explosion phase) and hence all the information regarding place of articulation is carried in the transition just before implosion.

Examples of the changes in vowel spectrum preceding the implosion phase of the voiced stop consonants /b/ and /d/ are shown in Fig. 12. Spectra sampled about 10 msec and 70 msec prior to consonantal closure are displayed. In the case of /b/ following the vowel /a/, there is little shift in F_2 during this interval, while F_1 falls slightly. The falling transition of F_2 can be seen clearly in the data for /d/. In both of these examples the consonant was exploded, and hence it is expected that additional information with

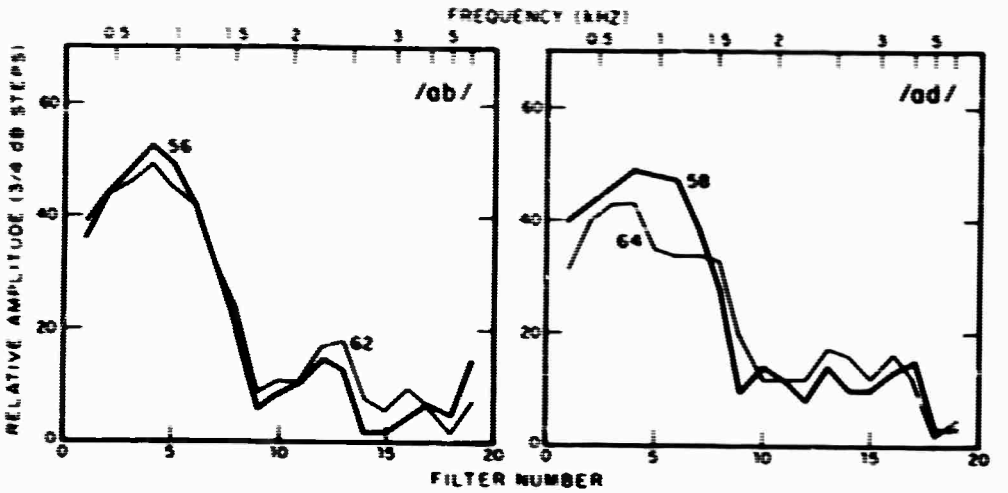


FIG.12 Spectra sampled 10 msec prior to consonantal closure (light lines) and 70 msec prior to consonantal closure (heavy lines) in syllables /ob/ and /od/. Speaker KS

regard to place of articulation is contained in the burst of energy following the consonantal release. In general, both aspects of a final stop consonant need to be examined and measured if the place of articulation is to be determined reliably.

3.3.3 Final sonorant (and nasal) consonants

The sonorant consonants that can occur in final position are /l/ and /r/ and the nasals /m, n, ŋ/. It is possible also to categorize /y/ and /w/ as consonants that can occur following stressed vowels, as in the diphthongs (such as /ɑl/, sometimes written /ɑy/) and the diphthongized vowels (such as /i, o/, sometimes written /iy, ow/). Examples of spectra sampled within the constricted interval in final position for the consonants /m, n, ŋ, l, r/ are shown in Fig. 13 for one speaker. All these spectra have very little energy above about 2000 Hz (filter 11); hence, they can be distinguished easily from the glide /y/ in a syllable like /bɑib/ (see SK I, Fig. 5, p. 22). The spectra for all the nasals are quite similar, with a low-frequency peak at filter 1 or filter 2 and with a rapid drop in spectrum at filters 3 and 4. The /l/ spectrum has a broad low-frequency peak extending over filters 2–4, corresponding to the closely spaced first- and second-formant frequencies. For /r/ the first-formant peak is at filter 2 and there is another important peak in the vicinity of 1500 Hz, corresponding to the combination of F_2 and F_3. These major spectral features appear also to characterize these final consonants for the other two speakers.

From data of this kind, it is probable that simple measurements on the spectra could be used to distinguish the nasals as a class, and could separate /l/ and /r/. Reference to the vowel spectra in

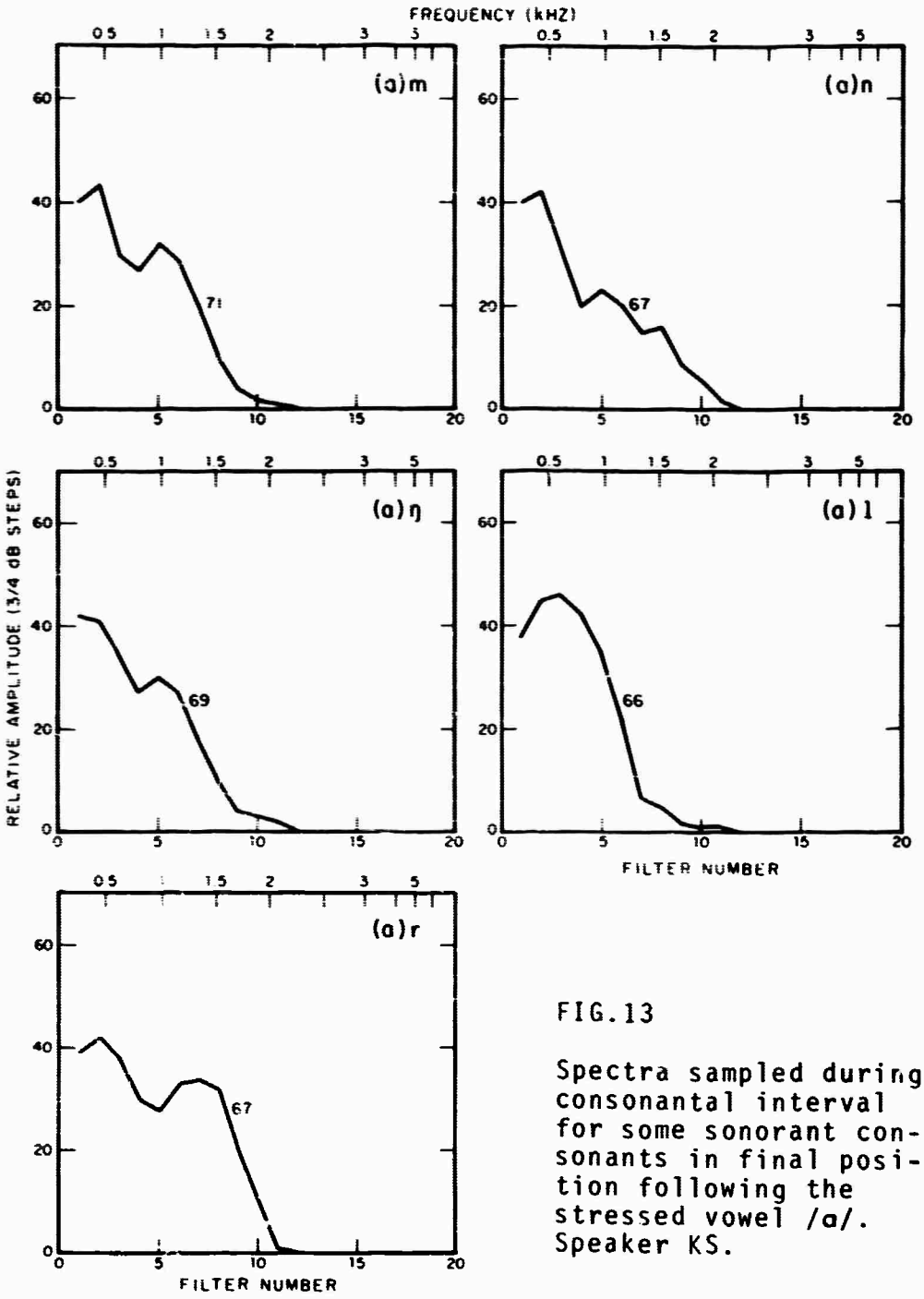


FIG.13

Spectra sampled during consonantal interval for some sonorant consonants in final position following the stressed vowel /ɑ/. Speaker KS.

SK I (see SK I, Fig. 5, p. 22) suggests that some difficulties might arise in separating /l/ from /w/ in final position using these kinds of spectral data. However, other information relating to the timing of the syllable could be used for these purposes. For example, *bow* (as in "bow and arrow") and *bowl* have quite different durations for the vowel and sonorant regions. Furthermore, pairs like *ball* and *bow* (as in "bow of a ship") are separable because the vowel qualities are different.

Some indication of the distinction between /m/, /n/, and /ŋ/ in final position can be seen in the graphs of Fig. 14. This figure shows spectra sampled about 20 msec prior to consonantal closure and again 30 msec after consonantal closure. The distinguishing attribute of the /n/ is the relatively sharp drop in the spectrum at high frequencies, particularly in the vicinity of the second or third formants. This drop occurs presumably because a zero is inserted in the vocal-tract transfer function in this frequency range at the instant when consonantal closure occurs. In the case of /ŋ/, an energy peak remains in the spectrum in the vicinity of the second formant of the vowel after consonantal closure occurs. The detailed characteristics of these nasal consonants may depend on the preceding vowel. In some cases, a vowel preceding a nasal consonant is nasalized, and the influence of this nasalization can be seen in the vowel spectrum. For the vowel /i/ in the syllable *in*, in Fig. 14, the nasalization is manifested as a bump in the spectrum in the vicinity of filter 5, at a point where a formant would not be expected for this vowel.

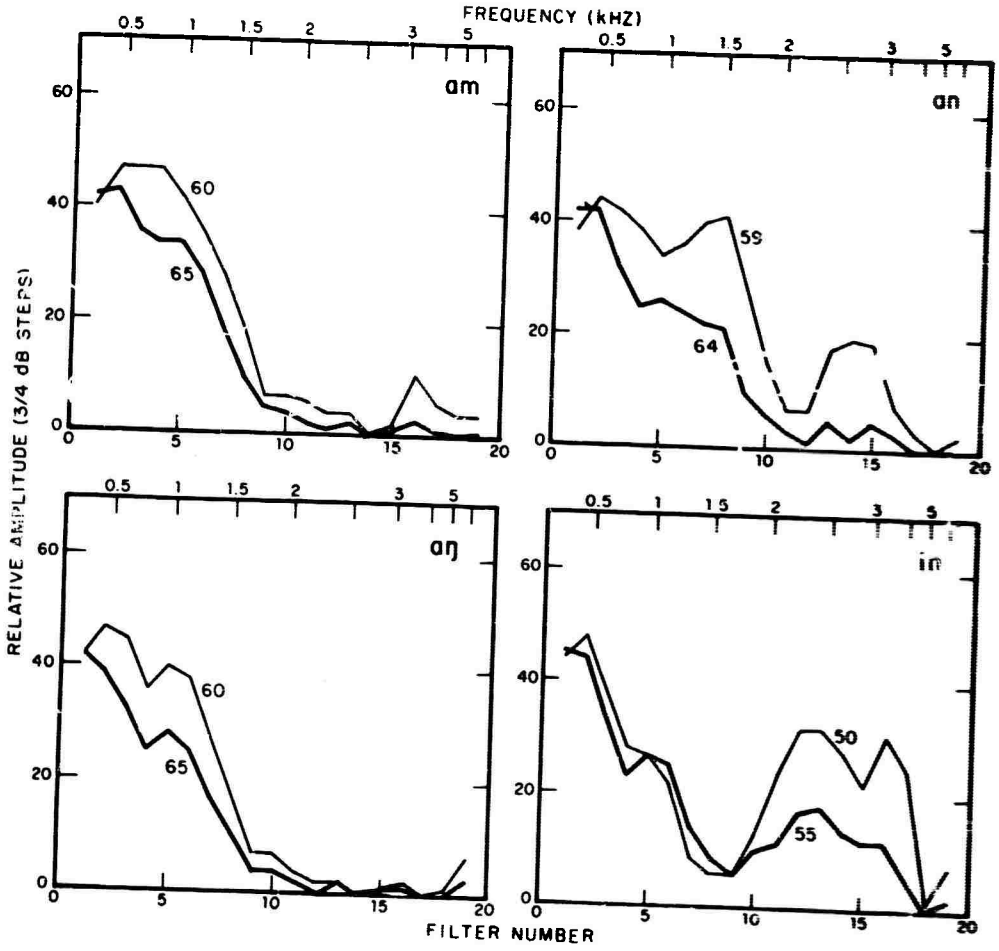


FIG.14 Spectra sampled 20 msec before consonantal closure (light lines) and 30 msec after closure (heavy lines) for nasal consonants in final position. Speaker KS.

4. REMARKS ON THE USE OF ACOUSTIC DATA IN SCHEMES FOR MACHINE RECOGNITION OF SPEECH

Approaches to automatic speech recognition have generally followed two different paths:

(1) Recognition is restricted to a closed set of utterances, each produced in isolation. A set of properties is extracted from the acoustic signal corresponding to each utterance, and a decision procedure operates on these properties to identify the utterance. It is necessary to include a learning phase in which distributions of the properties for each utterance (repeated many times) are determined. The decision procedures are based on observation of these distributions.

(2) The second approach involves no adaptive procedure or learning phase but recognizes that an utterance is constructed from a rather limited set of linguistic units. The recognition routine attempts to identify these smaller units (such as phonemes or phonetic features) and, from the results of these identifications, to recognize the entire utterances. In this situation, a lexicon that specifies the inventory of utterances in terms of the smaller phonetic units or features may or may not be a part of the overall recognition procedure, but usually it is a necessary component.

One basic problem in the first approach is the selection of properties sufficiently invariant from one repetition of an utterance to another. This selection is a most difficult task for several reasons. Perhaps the most important reason stems from the fact that, for any utterance in a language, there are rather severe constraints on the patterns of features that are allowed, i.e., on the phoneme classes and phoneme sequences that can occur in

the language. Because of these constraints, it is not necessary for the acoustic signal to carry precise information concerning every feature. A native speaker of the language somehow knows the rules (usually called phonological rules) governing the allowed patterns, and when he listens to another speaker, he can "fill in" information lacking in the signal by invoking these rules. Consequently, a speaker has a great deal of choice in how precisely he generates acoustic information concerning the various features. Often the acoustic information that corresponds to a given feature may be nonexistent or fragmentary, since the phonological rules (together with a lexicon) can specify the feature or at least can indicate that the feature is highly probable. It is quite possible, however, that a feature having weak or nonexistent acoustic correlates in one utterance may have important and essential acoustic correlates in another utterance of the same word or phrase.

Thus, for example, a speaker producing a simple word like *legal* has several possibilities open to him. The intervocalic /g/ may be produced as a velar fricative rather than as a stop. (This substitution would cause no confusion to an English listener who knows that a velar fricative is not allowed in his language and, therefore, must be interpreted as a stop.) The final unstressed syllable may be produced either as a schwa followed by /l/ or simply as a syllabic /l/. (No confusion arises here, since there can be no vowel contrast in this unstressed position.) Furthermore, the stressed vowel could be produced as /i/ or as /I/ or as anything in between without causing confusion. Thus there is a whole range of possibilities for generating the entire word.

A recognition scheme based on a learning or adaptive procedure would have to be exposed to all of these (and other) possibilities for the word during the learning phase, if it were to operate

satisfactorily for several speakers, or even for one speaker unless the speaker were carefully trained. It might be argued, of course, that as long as *some* of the properties of a given test utterance match, or are similar to, the set of stored properties for an item in the lexicon, then correct recognition might be achieved. That is, in a recognition task involving a limited vocabulary there might be sufficient redundancy that only certain attributes of an input utterance need to provide a match with the stored attributes. Except for situations involving a rather limited inventory of carefully selected utterances, however, this approach to speech recognition would have little chance of success.

The second approach to speech recognition (in which phonetic units or features are identified) also represents a most difficult problem, since it is necessary to store within the recognizer knowledge of the phonological rules that are possessed by a native speaker of the language. Also, and more important, the acoustic representation of a feature may vary tremendously depending on the environment of other features in which it occurs, as has been noted in this report and in SK I. The potential advantages of working toward a representation of an utterance in terms of features are (1) at least *some* features for some environments have well defined and reasonably stable acoustic correlates, and (2) a representation based on features is a convenient framework in terms of which the influence of environment on the acoustic correlates of a feature can be stated and the rules governing the allowed patterns of features in a language can be specified. Thus, in the example cited above, it would be easy to indicate in these terms that a velar fricative is not allowed (or alternatively, that recognition of the fact that a consonant is a velar automatically requires that it be a stop). It would also be a simple matter to require that all syllables with no stress consist of a vocalic schwa segment

that may or may not be followed by a consonant and to require that syllabic sonorant consonants be represented in these terms.

The point of this discussion is that a native speaker of a language has knowledge of the constraints on the possible patterns of features and feature sequences that can occur in an utterance. If the acoustic manifestation of a given feature is absent or distorted, this feature can be filled in by the listener on the basis of his knowledge of the constraints. The particular features that are distorted or missing in a given utterance may vary from one repetition of the utterance to the next. Some acoustic aspects of an utterance must, of course, provide clear and unequivocal information concerning the identity of a feature or features; the acoustic correlates of other features may be sufficiently distorted that they can only provide corroborative information when the context already permits strong hypotheses to be made concerning the identity of these features.

In view of these remarks, the acoustic data presented in this report and in SK I cannot be expected to encompass the properties of phonetic segments in all possible phonetic environments. An attempt is made to give examples of acoustic attributes in situations where they are reasonably unambiguous. Some data are presented, however, to indicate how these attributes may be modified in other phonetic environments. Furthermore, these acoustic data represent only a part of the knowledge that must be available to a machine that is to recognize speech. Also, the machine must be equipped with a set of rules specifying the constraints on patterns of features that are allowed in English and with a strategy for taking these rules into account.

BLANK PAGE

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Massachusetts 02138	Unclassified 2b. GROUP

3. REPORT TITLE

STUDY OF ACOUSTIC PROPERTIES OF SPEECH SOUNDS II, AND SOME REMARKS ON THE USE OF ACOUSTIC DATA IN SCHEMES FOR MACHINE RECOGNITION OF SPEECH

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific: Interim

5. AUTHOR(S) (First name, middle initial, last name)

Kenneth N. Stevens

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
15 August 1969	50	2

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F19628-68-C-0125/ARPA Order No. 627 b. PROJECT NO. 8668 c. DoD Element 6154501R d. DoD Subelement n/a	BBN Report No. 1871 Scientific Report No. 12 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFCRL-69-0339

10. DISTRIBUTION STATEMENT

Notice No. 1: Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
This research was sponsored by the Advanced Research Projects Agency under ARPA Order No. 627	Air Force Cambridge Research Labs (CRB) L.G. Hanscom Field Bedford, Massachusetts 01730

13. ABSTRACT

The acoustic properties of a number of different speech sounds as they appear in several phonetic contexts are described. This report supplements an earlier report on the same topic and presents data for stop and nasal consonants in prestressed position, for the timing of vowels, and for acoustic events following stressed vowels. The aims of this survey are to provide an indication of the kinds of accustic attributes that should be extracted from the speech signal in a potential scheme for machine recognition of speech. Also included is a discussion of the roles that must be played by acoustic data and by linguistic constraints in schemes for automatic speech recognition.

DD FORM 1473 (PAGE 1)

S/N 0101-807-6801

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech						
Phonetics						
Speech recognition						
Acoustic phonetics						